Submitter Name: Dajiang J. Liu
Submitted email: dajiang.liu@psu.edu

Phenotype imputation integrating GWAS summary association statistics, deeply phenotyped cohorts and large biobanks Identifies novel loci for cotinine levels

Lina Yang[1], Dajiang J. Liu[1]

[1]Department of public health sciences, Penn State College of Medicine

Large biobanks integrated with EMRs and omics data have greatly empowered genetic studies. Despite the benefits, biobanks often lack detailed measurements for studying behavioral traits, e.g. addictions, whereas detailed behavioral measures are often only available in smaller behavioral genetics studies. Phenotype imputation predicts unmeasured traits using estimated trait correlations, which has the potential of combining the strengths of biobanks and small behavioral studies. Existing methods such as MICE or SOFTIMPUTE rely solely on trait correlation, but ignore genetic data. A new method PHENIX integrates genetic data and explicitly models genetic correlation, but fails to integrate GWAS summary statistics from large datasets and is computationally intensive. We developed a novel phenotype imputation method based on Gaussian graphic models, which effectively and efficiently integrates GWAS summary statistics, biobanks with genetic but less detailed phenotypic data, and datasets with rich phenotypes. The new method greatly improves the estimates of genetic correlation and hence the imputation accuracy. Applying phenotype imputation, we imputed cotinine levels - a smoking biomarker, into UK Biobank (N=488,517) based upon the detailed phenotypes from the PASS smoking trial and GSCAN consortium GWAS summary statistics on common smoking traits. We identified 3 novel loci for cotinine including *DRD2* and *CHRNA5*, which greatly improved the power of an earlier study using only measured cotinine levels (N=4548) and imputation-based studies using alternative methods. The genetic correlation between imputed and measured cotinine levels are high (r2=.73). We expect our approach to be an extremely valuable tool to integrate small deeply-phenotyped studies with large biobanks.